Journal of Classification 19:249-276 (2002)
DOI: 10.1007/s00357-001-0045-7

Fixed Point Clusters for Linear Regression: Computation and Comparison

Christian Hennig

Universität Hamburg, ETH-Zürich

Abstract: A subset of a regression dataset, i.e., consisting of an independent variable and one or more regressors, is called regression Fixed Point Cluster (FPC) if it reproduces itself under the following procedure: Its linear regression and variance estimators are computed, all points too far from the regression hyperplane are declared as outliers, and the subset under consideration is exactly the set of non-outliers w.r.t. itself.

In this paper an algorithm is developed, which aims to find all FPCs of a dataset corresponding to well separated linear regression subpopulations. Its ability to find such subpopulations under the occurrence of outliers is compared to methods based on ML-estimation of mixture models by means of a simulation study. Furthermore, FPC analysis is applied to a real dataset.

Keywords: Switching regression; Robust regression; Outlier identification; Mixture model; Gaussian mixtures with noise; Clusterwise linear regression.

Author's Address: Christian Hennig, Fachbereich Mathematik, Universität Hamburg and (currently) Seminar für Statistik, ETH-Zentrum, CH-8092 Zürich, Switzerland; telephone: +41 1 632 6184; fax: +41 1 632 1228; e-mail: hennig@math.uni-hamburg.de.

1. Introduction

Cluster analysis is related to the concept of outliers. If a part of a dataset forms a well separated cluster, this means that the other points of the dataset appear outlying with respect to the cluster. It may be interpreted synonymously that "a cluster is homogeneous" and that "it does not contain any outlier". The idea of Fixed Point Cluster (FPC) analysis (Hennig 1997, 1998) is to formalize a cluster as a data subset that does not contain any outlier and with respect to that all other data points are outliers.

The concept is applied to clusterwise linear regression in this paper. That is, a relation

$$y = \mathbf{x}'\beta + u,\ E(u) = 0, \tag{1}$$

between a dependent variable y and an independent variable $\mathbf{x} \in I\!R^p \times \{1\}$ (β_{p+1} denoting the intercept parameter) should be adequate for a single cluster. The values of $\mathbf{x}$ can be fixed or random. The error variable u is assumed as Gaussian distributed with variance σ^2 independent of $\mathbf{x}$. There are lots of applications of clusterwise linear regression, e.g. in biology (Hosmer 1974) and economics (DeSarbo and Cron 1988, Wedel and DeSarbo 1995, Wedel and Kamakura 2000). A further example is given in Section 5.

Most of the methods for clusterwise linear regression are based on least squares and maximum likelihood estimation for mixture and partition models. For overviews see Wedel and Kamakura (2000), Hennig (1999). Viele and Tong (2000) treat such mixtures from a Bayesian viewpoint, Cook and Critchley (2000) develop a method to detect regression clusters graphically, Shannon, Faifer, Province and Rao (2002) use a tree-based approach, and Morgenthaler (1990) uses the local minima of redescending M-estimators to find such clusters. Morgenthaler's approach is an ancestor of FPC analysis, as explained in Section 2.

To illustrate the idea of FPC analysis, consider the artificial dataset of Figure 1. From the viewpoint of robust statistics, the dataset can be interpreted as a main part of 90 points (circles), Gaussian distributed along a linear regression line, and 11 outliers. Davies and Gather (1993) emphasize that the term "outlier" should be defined with respect to a reference distribution. That is, the shape of the "good" points has to be assumed to define what a "bad" point is, and the region of outliers can be defined e.g. as a region where the density of the reference distribution is low. In this sense, the 11 points indicated by triangles and diamonds are outliers with respect to the distribution underlying the data subset of circles, and the circles are the set of non-outliers with respect to themselves. However, while it is usually assumed in robust statistics that at least half of the points are non-outliers, we observe that the 10 points indicated by triangles can as well be considered as non-outliers with respect to their own

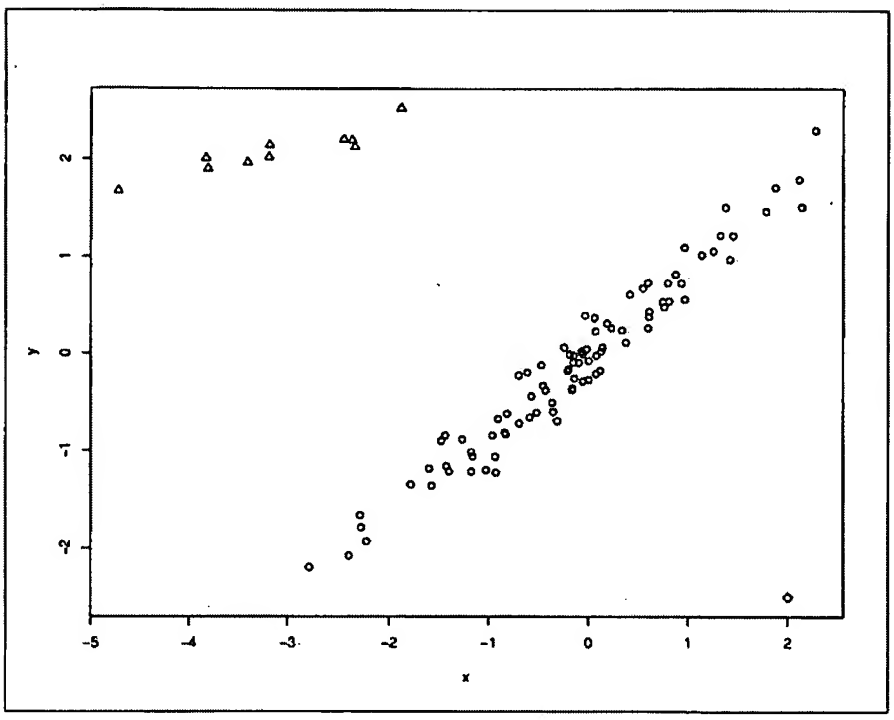

Figure 1. Artificial data.

underlying linear regression distribution, in which case all remaining 91 points are defined as outliers.

Such subsets are in some sense homogeneous (they do not contain any outlier) and well separated (all other points are outliers with respect to them) and therefore it is reasonable to call them "clusters".

Note that not all data subsets are exactly the non-outliers with respect to themselves: If the distributions of the type (1) without further assumptions to the x-values are chosen as the class of reference distributions, the 11 non-circles in Figure 1 cannot all at the same time be reasonably defined as non-outliers unless at least some of the circles in between are non-outliers, too.

"Consisting of the non-outliers with respect to itself" can be formalized as a fixed point condition, which is done in Section 2. For combinatorial reasons, it is impossible to check this condition for all data subsets. Therefore, an implementation is needed to find with high probability all meaningful fixed point clusters of a dataset. Such an implementation is given in Section 3. In Section 4, FPC analysis is compared by means of a simulation study to an alternative procedure for clusterwise linear regression, namely the mixture model

ML-estimator introduced by DeSarbo and Cron (1988). Apart from FPCs, I do not know of any clusterwise linear regression procedure that accounts explicitly for outliers, although the approach of using mixtures of t-distributions (Peel and McLachlan 2000) may be easily extended to the linear regression setup. If the distribution of the independent variable is not too far from a Gaussian distribution, it is possible to apply the MCLUST-package discussed by Fraley and Raftery (1998, 1999), which performs an ML-estimation of a mixture model of multivariate Gaussian distributions and allows for noise modeled by a Poisson process component. The package is suggested for "linearly shaped clusters" by DasGupta and Raftery (1998), and I included it in the simulation study (it is not claimed that FPC analysis as defined here is a competitor with MCLUST in non-regression setups). Other packages for Gaussian mixtures such as EMMIX by McLachlan et al. (1999), which also fits more robust ML-estimators for mixtures of t-distributions, could be used as well. It should be emphasized, however, that in clusterwise linear regression data subsets are interpreted as homogeneous groups if they follow a common linear relation between the independent and the dependent variable, while clustering by Gaussian or t-mixtures attempts to divide such groups if the independent variable alone is non-homogeneous.

FPC analysis is not as good as the ML methods if their model assumptions are fulfilled exactly, but it turns out to be clearly superior to regression mixture ML under the occurrence of outliers and to MCLUST under strongly non-Gaussian independent variables. Furthermore, overlapping FPCs may give some additional insights in the data structure. In Section 5, FPC analysis is applied to a dataset concerning tone perception.

While the simulation study treats the recovery of Gaussian mixture components under the occurrence of outliers, FPC analysis is essentially not a method to estimate mixture components. FPCs can be interpreted as estimators for theoretical FPCs of populations (as defined in Section 2), which provide an alternative *definition* of a cluster in a stochastic model. Theoretical FPCs are treated elsewhere in greater detail (Hennig 2000a), as well as the consistency of data FPCs for them (Hennig 2000c).

2. Clusters and Outliers in Linear Regression

An outlier identification procedure is needed to formalize the idea outlined in the introduction. For a reference model $P_{\beta,\sigma^2,G}$ of the form (1), G denoting a regressor distribution, the easiest outlier region in the spirit of Davies and Gather (1993) is defined as

$$A(c, P_{\beta,\sigma^2,G}) := \{(\mathbf{x}, y) \in I\!R^p \times \{1\} \times I\!R : (y - \mathbf{x}'\beta)^2 > c\sigma^2\}, \quad (2)$$

that is, points outside a strip of width $2\sqrt{c\sigma^2}$ around the regression hyperplane defined by β are defined as outliers. The tuning constant c can be chosen as a large quantile of the χ^2_1-distribution to get a low probability for the occurrence of outliers under $P_{\beta,\sigma^2,G}$.

This definition is used to define an FPC with respect to a distribution P on $\mathbb{R}^{p+1}$: An FPC should be a set S which contains exactly the non-outliers with respect to itself. For a given distribution P and set S let P_S denote the conditional distribution of P on S. For arbitrary P let $\beta(P)$ and $\sigma^2(P)$ denote the LS-linear regression and error variance functional, respectively ($A = \emptyset$ in case of non-existence).

For $(\beta,\sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}_0^+$ let $h(\beta,\sigma^2) = A(c, P_{\beta,\sigma^2,G})^c$ from (2) be the corresponding region of non-outliers, which does not depend on G. Now a set S containing exactly all non-outliers with respect to itself, i.e., all points close enough to the regression hyperplane defined by the parameters of P_S, is defined as follows:

Definition 2.1 *A set S fulfilling $S = f(S)$, $f(S) = h(g(S))$, $g(S) = (\beta(P_S), \sigma^2(P_S))$ is called a theoretical FPC of a distribution P.*

This could be written as well as a fixed point condition concerning (β,σ^2), namely by considering $h \circ g$. Examples of theoretical FPCs are given in Hennig (2000a).

FPCs for datasets can be defined by replacing $\beta(P)$ and $\sigma^2(P)$ by estimators. I introduce some notation first: Let $(\mathbf{X},\mathbf{y}) := ((\mathbf{x}_1',y_1),\ldots,(\mathbf{x}_n',y_n))'$, where $\mathbf{x}_i \in \mathbb{R}^p \times \{1\}$, $y_i \in \mathbb{R}$, $i = 1,\ldots,n$, be a regression dataset. Let $\mathbf{W}_{\beta,\sigma^2}$ be the diagonal matrix of indicators (weights)

$$1[(y_i - \mathbf{x}_i'\beta)^2 \leq c\sigma^2],\ i = 1,\ldots,n,$$

of the points lying close enough to the regression hyperplane defined by β. Let $n(\mathbf{W}) = \text{trace}(\mathbf{W})$ the number of points indicated by $\mathbf{W}$.

Definition 2.2 *A diagonal matrix $\mathbf{W}_{\beta,\sigma^2}$ with diagonal in $\{0,1\}^n$ is called* **Fixed Point Cluster Matrix** *(FPCM) w.r.t $(\mathbf{X},\mathbf{y})$ (and the indicated points form an FPC), iff $(\beta,\sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}_0^+$ is a fixed point of*

$$f:\ (\beta,\sigma^2) \mapsto \left(\hat{\beta}\left(\mathbf{W}_{\beta,\sigma^2}\right), \hat{\sigma}^2\left(\mathbf{W}_{\beta,\sigma^2}\right)\right),$$

$$\text{where } \hat{\beta}(\mathbf{W}) := (\mathbf{X}'\mathbf{W}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y},$$

$$\hat{\sigma}^2(\mathbf{W}) := \tfrac{1}{n(\mathbf{W})-p-1}\left(\mathbf{y} - \mathbf{X}\hat{\beta}(\mathbf{W})\right)'\mathbf{W}\left(\mathbf{y} - \mathbf{X}\hat{\beta}(\mathbf{W})\right).$$

In case of the non-existence of $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$, $f(\beta,\sigma^2) := (\beta,\infty)$.

Analogously to Definition 2.1, $\mathbf{W}$ is required to indicate exactly the points close enough (in terms of the error variance estimator $\hat{\sigma}^2(\mathbf{W})$) to the regression hyperplane defined by the regression estimator $\hat{\beta}(\mathbf{W})$. That is, to verify the FPC property of a data subset, its regression and error variance estimator have to be calculated and it has to be checked if all its points (and no others) are inside the corresponding strip around the regression hyperplane. This holds for the subset of circles as well as for the subset of triangles in Figure 1 if c is chosen suitably, e.g. $c = 6.635$, being the 0.99-quantile of the χ^2-distribution. The choice of c is discussed in the Sections 3.5 and 4.1. Note that FPCs inherit the equivariance properties of the LS-regression and scale estimator.

Most recent outlier identifiers are based on robust estimators, see e.g. Davies and Gather (1993), Rocke and Woodruff (1996) and Hawkins (1999), and it may be wondered why the outlier identifier used in Definition 2.2 is based on the LS-estimator and the corresponding error variance. Indeed, it would be of interest to use a more sophisticated outlier identifier to define FPCs, but this leads to some additional theoretical and computational difficulties and is left to future research. On the other hand, Kosinski (1999) demonstrates that an outlier identifier based on calculation of non-robust mean, variance and covariance estimators on data subsets can have a good performance as well.

The definition of FPCs given here is a further development from the following approach of Morgenthaler (1990), which is related to robust methods: A redescending M-estimator is defined by

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i'\beta}{s}\right), \tag{3}$$

where ρ is symmetric about 0, monotone increasing between 0 and some $k > 0$ and constant outside $(-k, k)$. s is an error scale constant and has to be chosen in advance. A local minimum of (3) fulfills the fixed point condition

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}(\hat{\beta})\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}(\hat{\beta})\mathbf{y}, \tag{4}$$

where $\mathbf{W}(\hat{\beta})$ is a diagonal matrix of weights $w_i(\beta) = \frac{\psi((y_i - \mathbf{x}_i'\beta)/s)}{(y_i - \mathbf{x}_i'\beta)/s}$, $\psi = \rho'$, possibly piecewise (Huber 1981, p. 183 ff.). That is, such a local minimum can be interpreted as a weighted LS-estimator, where observations with $(y_i - \mathbf{x}_i'\beta)^2/s^2 > k^2$ get a weight of zero and can be interpreted as outliers, because of $\psi = 0$ outside $(-k, k)$. This is why these estimators are called "redescending".

Redescending M-estimators are not so popular in robust statistics, because the non-uniqueness of the solutions of (4) poses certain theoretical and practical problems. However, Morgenthaler (1990) utilized this non-uniqueness to locate different subpopulations (clusters). Such multiple solutions can be

found by use of a fixed point algorithm (sometimes called "iteratively reweighted least squares" in this setup) from varying starting values. The general idea to use solutions of the fixed point equation of redescending M-estimators for clustering goes back to Hampel (1975).

FPC analysis as defined here uses the same idea with binary weights (more general weights would be possible) and defines the error scale by a fixed point equation analogous to (4) to account for clusters with differing error scales. Therefore, its solutions are no longer local minima of a global objective function such as (3). There is no ordering between the clusters.

Performing a cluster analysis in such a way has two main advantages:

1. The number of clusters does not need to be specified in advance.

2. The parameter estimator for one cluster is not only unaffected by points from the other clusters (as long as they are well separated), but furthermore the influence of observations which lie far from *any* linear regression population is weighted down to zero. That is, not all points (and not even their majority) need to belong to linear regression clusters. This is different from clustering based on mixture models: While it is possible for methods like those of DasGupta and Raftery (1998) and Peel and McLachlan (2000) to handle easy situations like the diamond outlier of Figure 1 adequately, it cannot be ruled out that the addition of further outliers affects the parameter estimators for the points belonging to homogeneous subpopulations.

A third difference to methods based on mixture or partition models is that observations may belong to more than one cluster, which may be seen as an advantage or as a drawback, see the discussion in Section 5. (The a posteriori membership probabilities from mixture model ML-estimators can be larger than zero for more than one cluster, but nevertheless each observation is modeled as coming from exactly one component in a mixture model.) In particular, all subsets lying exactly on a regression hyperplane, i.e., $\sigma^2 = 0$, form FPCs unless the covariate points are collinear. For very small subsets, this may not be very interesting, but larger exactly linear portions of the data are often meaningful. It will be discussed later how to avoid meaningless FPCs in the output of the procedure. Note that the system of FPCs does not have a hierarchical structure in general.

It follows directly from Definition 2.2 that outliers outside the non-outlier strip of an FPC never destroy its FPC property. This may be seen as a robustness property of FPC analysis. Unfortunately, it cannot be formalized in terms of a high breakdown point, because under certain circumstances, robustness problems can be caused by points *inside*, but near to the border of the strip. Note, however, that formal robustness properties even of the simplest cluster analysis

procedures are difficult to obtain, see Garcia-Escudero and Gordaliza (1999), Kharin (1996, Section 7.6).

The relation between clustering and outliers was considered recently as useful for the sake of outlier identification as well, see Rocke and Woodruff (1996, 2001).

3. Implementation

3.1 A Fixed Point Algorithm

It is practically impossible to check the FPC-property of every subset of a dataset, except if it is very small. But FPCMs can be found by means of a fixed point algorithm. Its convergence is shown as Theorem 3.1.

Fixed point algorithm (FPA): Choose a diagonal indicator matrix $\mathbf{W}^0$ with $n(\mathbf{W}^0) > p+1$, $k=0$.

Step 1 Compute $\hat{\beta}(\mathbf{W}^k), \hat{\sigma}^2(\mathbf{W}^k)$.

Step 2 $\mathbf{W}^{k+1} = \text{diag}\left(1[(y_i - \mathbf{x}_i'\hat{\beta}(\mathbf{W}^k))^2 \leq c\sigma^2(\mathbf{W}^k)]\right)$.

Step 3 End if $\mathbf{W}^k = \mathbf{W}^{k+1}$, else $k = k+1$, step 1.

Theorem 3.1 *Let $c > 1$. If $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ exists whenever $n(\mathbf{W}) > p+1$, then for some $k < \infty$: $\mathbf{W}^k = \mathbf{W}^{k+1}$, i.e., the FPA converges in finitely many steps.*

The proof is given in the appendix.

3.2 Outline of the Implementation

Because more than one FPC is of interest in cluster analysis, it does not suffice to run the FPA only once. The rest of Section 3 is devoted to the development of a procedure which should be able to discover all substantial FPCs with high probability, where "substantial" means all FPCs corresponding to well separated, not too small data subsets which can adequately be described by linear regression distributions of type (1). The implementation involves the choice of some constants, which is discussed in detail. A short description of the implementation is given in Section 3.6. The basic procedure is simple:

1. Choose the number of algorithm runs $i_{n,p}$ and the tuning constant c.
2. Repeat $i_{n,p}$ times: Generate an indicator matrix $\mathbf{W}^0$ with $n(\mathbf{W}^0) = p+2$ randomly and apply the FPA until convergence. The choice of $n(\mathbf{W}^0)$ is justified in Section 3.3. Store all found FPCMs and count the number of times that each FPCM has been found.

The choice of $i_{n,p}$ and c is discussed in the Sections 3.3 and 3.5.

Applying the basic procedure, one may observe that the number of found FPCs is often larger than one would like to have for the clarity of the interpretation, unless n is very large or $i_{n,p}$ is so small that the result of the analysis depends strongly upon chance. There are some reasons for that:

- Often more than one version of the same visible cluster in the data manifests itself as an FPCM, because FPC analysis allows an arbitrarily large overlap between the clusters.
- Often there appear small subsets of a dataset that can be fitted almost exactly by a regression hyperplane. Because their separation from the remaining points is measured by their very small error variance, they can meet the FPCM definition. If $i_{n,p}$ is large enough, some algorithm runs lead to very small FPCMs by chance.
- If the dataset is sparse in $\mathbb{R}^{p+1}$, i.e., if n is small or p is large, it is not very difficult for subsets to be separated well enough from the rest of the data and thus to be FPCMs. Partitioning methods and methods based on mixture models suffer from lots of local optima of the criterion function in this situation, while FPC analysis yields lots of FPCs.

An applicable procedure needs a third step, which will be described precisely in the Sections 3.4 and 3.6:

3. Reduce the number of FPCs to be interpreted by
   - defining groups of similar found FPCMs (*clusters of clusters*),
   - discarding all FPCMs of groups that are found too seldom,
   - choosing a representative FPCM for each of the remaining groups.

3.3 The Number of Algorithm Runs

Suppose that a dataset of size n contains an FPCM $\mathbf{W}$ of size $n(\mathbf{W})$, which is homogeneous and well separated from the rest of the data. The number of algorithm runs should be chosen in order to find such an FPCM with high probability. This probability can be calculated on the base of the probability $p_{\mathbf{W}}$ that this FPC is found by a single run of the FPA, started with $n(\mathbf{W}^0)$ randomly selected points. Of course, $p_{\mathbf{W}}$ cannot be determined exactly. However, when the data can be partitioned in "good points" and outliers, subset based methods of estimation and outlier identification behave robustly if they start with a subset consisting only of good observations (see e.g. Rousseeuw and Leroy 1987, Kosinski 1999). If the points of $\mathbf{W}$ are considered as "good", a

Table 1: Number of algorithm runs $i_{n,p}$

n	p	$i_{\min}$	$i_{n,p}$	n	p	$i_{\min}$	$i_{n,p}$
50	1	1	487	50	1	3	1027
200	1	1	396	200	1	3	834
50	2	1	3283	50	2	3	6903
200	2	1	2119	200	2	3	4453
200	4	1	64316	200	4	3	135167
1000	4	1	49735	1000	4	3	104523

starting configuration with points only from $\mathbf{W}$ will lead to a relatively reliable non-outlier region. The FPA starting from such points will usually lead to $\mathbf{W}$ or a very similar FPC (which may exist), if $\mathbf{W}$ is well separated enough and does not have significantly separated subpopulations (in which case these subpopulations would be of main interest). Consequently, under the circumstances mentioned above, at least the probability of finding a member of the Single Linkage cluster of similar FPCs of $\mathbf{W}$ (see Section 3.4) can be approximated reasonably by

$$q_{n(\mathbf{W})} := \frac{\binom{n(\mathbf{W})}{n(\mathbf{W}^0)}}{\binom{n}{n(\mathbf{W}^0)}}. \tag{5}$$

The minimal possible $n(\mathbf{W}^0) = p + 2$ maximizes $q_{n(\mathbf{W})}$ and is therefore recommended. It cannot be expected to find FPCs with too small $n(\mathbf{W})/n$ in a feasible number of algorithm runs. Thus, a decision is necessary about the smallest size $n_{\min}$ of an FPCM that one wants to find with high probability.

$i_{n,p}$ can be chosen so that the approximated probability to start with $p+2$ points from an FPC with $n_{\min}$ points at least $i_{\min}$ times is at least 0.95:

$$i_{n,p} := \min\{i : \mathrm{QB}(i, q_{n_{\min}}; 0.05) < i_{\min}\}, \tag{6}$$

where $\mathrm{QB}(n, p; \alpha)$ denotes the α-quantile of the Binomial(n, p)-distribution (compare Appendix B of Kosinski 1999).

While $i_{\min} = 1$ minimizes $i_{n,p}$, I suggest to take a larger value of $i_{\min}$, say $i_{\min} = 3$, as long as the computation time allows, because this makes the result of the analysis more stable. Table 1 gives some values of $i_{n,p}$ for $n_{\min} = \frac{n}{5}$. Unfortunately, $i_{n,p}$ increases exponentially in p. It decreases moderately in n. The computation time for a single algorithm run increases in p and n, but apparently not very fast. However, $p > 4$ is not feasible at the moment, except with manually chosen cluster candidates as starting configurations.

3.4 Reduction of the Number of FPCs

Usually some of the $i_{n,p}$ algorithm runs lead to FPCs which are very similar to others or not very stable. The number of FPCs can be reduced by

defining representative FPCs for groups of similar FPCs, and by discarding apparently unstable FPCs.

The agglomeration of groups of FPCs requires a similarity measure between FPCs. Similarity between FPCs should be defined by means of the number of common data points and not by means of their regression and error variance parameters to keep the equivariance properties of FPCs. A similarity measure between the indicator vectors $\mathbf{v}$ and $\mathbf{w}$ of subsets of a dataset (main diagonals of indicator matrices $\mathbf{V}, \mathbf{W}$, respectively) is defined by relating the number of points of the intersection of the subsets to the sum of their sizes:

$$s_*(\mathbf{v}, \mathbf{w}) := \frac{2\mathbf{v}'\mathbf{w}}{\mathbf{v}'\mathbf{v} + \mathbf{w}'\mathbf{w}}, \tag{7}$$

so that $0 \leq s_*(\mathbf{v}, \mathbf{w}) \leq 1$, where $s_*(\mathbf{v}, \mathbf{w}) = 0$ iff $\mathbf{v}$ and $\mathbf{w}$ have no point in common, and $s_*(\mathbf{v}, \mathbf{w}) = 1$ iff $\mathbf{v} = \mathbf{w}$. To define a partition of the FPCMs, one can specify $0 < s_{cut} < 1$ so that $\mathbf{v}, \mathbf{w}$ are interpreted as "similar" if $s_*(\mathbf{v}, \mathbf{w}) \geq s_{cut}$. The Single Linkage clusters of index s_{cut} are defined as the connectivity components of the graph with the FPCMs as vertices and edges between all pairs $\mathbf{v}, \mathbf{w}$ where $s_*(\mathbf{v}, \mathbf{w}) \geq s_{cut}$. Cormen, Leiserson and Rivest (1990, p. 477) give an algorithm to compute these connectivity components without calculating the whole Single Linkage tree. This seems to be the easiest method to get a reasonable partition based on similarities without assumptions about the number of groups. I suppose $s_{cut} = 0.85$, which means that two FPCs of 20 points each are considered as similar if they have at least 17 points in common. A subset of at least 16 points is considered as similar to a set of 20 points.

Complete Linkage clustering could as well be reasonable, but if its cluster index s_{cut} is chosen moderately smaller than for the corresponding Single Linkage solution, it does not seem to lead to considerably different results. Chaining problems seem unlikely in this setup.

For each of the groups, a representative FPCM is chosen. The ratio of the number of findings $i_{\mathbf{W}}$ to its estimated expected value (based on the approximation $q_{n(\mathbf{W})}$)

$$r_{\mathbf{W}} = \frac{i_{\mathbf{W}}}{i_{n,p}q_{n(\mathbf{W})}} \tag{8}$$

measures the stability of an FPCM. Therefore, it is reasonable to choose the FPCM with the largest $r_{\mathbf{W}}$ as the representative FPCM.

The effect of the Single Linkage reduction can be seen in Table 2 in Section 4.1. However, my experience is that the effect of this reduction is more important in real data, where the tails of the error distributions are not exactly Gaussian and FPCs differ often by the inclusion or exclusion of single points in the error tail regions, compare the example in Section 5.

The similarity s_* is used in Section 4.2 as well to measure the quality of the cluster recovery by the compared cluster analysis methods.

Some FPCs can be considered as "too unstable", namely too small ones and the ones found too seldom. If $n(\mathbf{W}) < n_-$ defined by

$$n_- := \min\{\bar{n} : \ \mathrm{QB}(i_{n,p}, q_{\bar{n}}; 0.5) < i_{\min}\} \tag{9}$$

then an FPCM $\mathbf{W}$ will be reproduced in a further analysis with an estimated probability of at most 0.5. Thus it should be excluded. This should be done before the Single Linkage agglomeration to prevent that a whole group of larger FPCs is lost because of a small representative FPC that occurred often by chance.

If an FPCM is found seldom, there can be similar FPCMs corresponding to the same, possibly relevant, pattern of the data. But Single Linkage groups of FPCMs should be excluded as well, if their members are found less than i_{min} times.

3.5 The Tuning Constant c

The tuning constant c defines the size of the distance that a point must have from the center of a Gaussian distribution to call it an "outlier". The larger c, the more separation is needed for a data subset to meet the definition of FPCMs. Formally this can be seen by considering theoretical FPCs as defined in Definition 2.1. It can be shown (Hennig 2000c) that a single linear regression distribution of type (1) has a unique theoretical FPC for $c \geq 3$, which contains 98.5% of the mass for $c = \chi^2_{1;1-\alpha} = 6.635$ resulting in an error variance of 0.9001, more for larger c and less for smaller c. If a mixture of such a distribution P with another distribution Q is considered, it depends on the amount of mass from Q inside of $A(c, P_S)^c$, if there remains a theoretical FPC corresponding to P. The smaller c, the closer to P the regions of high density of Q may lie, so that there is still an FPC corresponding to P. Examples are given in Hennig (2000a).

But these are only asymptotical considerations, and there is an important effect of the sample size n in dependence of the dimension p. If n is small, the data can be so sparse that almost all strips around arbitrary regression hyperplanes contain data subsets that form FPCs because they are well separated enough from the rest of the data. That is, the larger p and the smaller n, the larger c is required to prevent the occurrence of such meaningless FPCs. This corresponds to Rousseeuw's (1994) words: *"My interpretation of the "curse of dimensionality" is that several structures can exist simultaneously in the same dataset."* Unfortunately, up to now there is no well-founded theory to relate n, p and c, and furthermore such a relation depends upon the other parameters

of the algorithm to find FPCs, which were discussed in the previous sections.

The approach taken here is to simulate the number of found FPCs of a homogeneous Gaussian linear regression population for the proposed values of n_{min}, i_{min} and s_{cut} and various values of n, p and c. The results are given in Section 4.1, Table 2. Formula (10) can be used to choose a value of c for particular values of n and p. It approximates roughly the value of c from the simulations so that it is as small as possible with sufficient certainty that only one FPC is found for a homogeneous population. More formally: In 20 simulation runs there should be at most a mean of 1.2 found FPCs, from which always only one is representative:

$$c(n,p) = 3 + \frac{33}{[2^{-(p-1)/2}n]^{1/3}} + \frac{2900000}{[2^{-(p-1)/2}n]^3}. \tag{10}$$

Table 2 shows the $c(n,p)$-values corresponding to n and p chosen for the simulation. If the asymptotical value c_a of c is desired to be larger than 3, $c = \max[c(n,p), c_a]$ can be chosen.

For small n, (10) leads to enormous values for c. It is not reasonable to demand a separation defined by more than, say, $c = 70$, for a clearly separated cluster, and therefore higher values of $c(n,p)$ indicate that data of the corresponding n and p are typically too sparse to apply FPC analysis reasonably. However, I made good experiences with the application of (10) for larger values of n than those chosen in the simulations.

3.6 A Description "Ready to Run"

Step 1 Choose c according to (10), $n_{\min} = \frac{n}{5}$, $i_{\min} = 3$, $s_{cut} = 0.85$. Of course, all these choices are subjective because they concern trade-offs between more information and better interpretability, more stability and lower computing time, respectively, as discussed in the previous sections.

Step 2 Compute $i_{n,p}$ according to (6), n_- according to (9).

Step 3 Repeat $i_{n,p}$ times: Generate an indicator matrix $\mathbf{W}^0$ with $n(\mathbf{W}^0) = p + 2$ by random and apply the FPA. Store all found FPCMs $\mathbf{W}$ with $n(\mathbf{W}) \geq n_-$. Count the number of times that each FPCM has been found.

Step 4 Compute the similarities for each pair of FPCMs according to (7).

Step 5 Compute the Single Linkage clusters of index s_{cut} of FPCMs.

Step 6 For $j = 1, \ldots, j_G$:

$$i_j := \sum_{\text{group}(\mathbf{W})=j} i_{\mathbf{W}}.$$

If $i_j \geq i_{\min}$, choose

$$\mathbf{W}_j^* := \underset{\text{group}(\mathbf{W})=j}{\arg\max} \{r_{\mathbf{W}}\},$$

$r_{\mathbf{W}}$ defined according to (8).

The $\mathbf{W}_j^*$, $j = 1, \ldots, j_G$, with $i_j \geq i_{\min}$ are the representative FPCMs. Let n_r denote its number in the following.

It might be disappointing that the result of an FPC analysis depends upon chance via the random starting configurations, that the approximations of the probability to find a relevant FPC are rough, and that the computing time gets very large for large p. However, minimization of multidimensional criterion functions as necessary in cluster analysis based on estimators in mixture models leads to similar problems. There are some reasonable approaches to use only one good starting point instead of many random configurations (e.g. De Veaux 1989, Fraley and Raftery 1999, Coleman and Woodruff 2000), but there is even less theoretical foundation in terms of the probability to discover a given cluster, and for large p to my knowledge there are also no simulations.

4. Simulations

4.1 FPCs in Homogeneous Populations and Choice of c

In a population consisting of only one homogeneous regression component asymptotically only one FPC is to be expected. The simulations of this section are to show the relation between the number of found FPCs, n, p, and the tuning constant c. Here, all points $(x_1, \ldots, x_p, y) \in I\!R^{p+1}$ were generated according to a $p+1$-dimensional Gaussian distribution with mean 0 and covariance matrix $\mathbf{I}_{p+1}$. The procedure of Section 3.6 was used.

There were 20 simulations runs for each constellation. The number of found FPCs n_c was recorded as well as the number of their Single Linkage clusters found often enough (n_r). The simulation results are given in Table 2.

While it is obvious that the number of FPCs decreases with increasing n, increasing c and decreasing p, these relations do not seem to follow a simple functional pattern. The formula (10) was fitted by "trial and error" to get not too large values of $c(n,p)$ which guarantee for not more than one cluster in a homogeneous population with high probability. The values of $c(n,p)$ for the simulated values of n and p are given as well in Table 2.

Note that because of the equivariance properties, the results generalize to arbitrary homogeneous linear regressions with Gaussian distributed independent variables. Unfortunately, arbitrarily distributed independent variables are not covered in general.

Table 2: Average number of representative FPCs (found FPCs) $n_r(n_c)$ for homogeneous data

$p=0$	$c=6$	$c=10$	$c=20$	$c=30$	$c(n,p)$
$n=25$	3.8 (9.4)	2.5 (3.9)	1.5 (2.4)	1.2 (1.5)	78.7
$n=50$	2.7 (6.8)	1.3 (2.1)	1.0 (1.1)	1.0 (1.0)	19.2
$n=100$	1.4 (2.6)	1.0 (1.1)	1.0 (1.0)	1.0 (1.0)	10.4
$n=200$	1.0 (2.2)	1.0 (1.1)	1.0 (1.0)	1.0 (1.0)	8.2
$p=1$	$c=6$	$c=10$	$c=20$	$c=30$	$c(n,p)$
$n=25$	15.7 (39.7)	9.4 (15.3)	3.9 (5.2)	2.6 (2.9)	199.9
$n=50$	9.6 (39.5)	3.2 (10.4)	1.4 (2.7)	1.0 (1.5)	35.2
$n=100$	2.6 (9.7)	1.1 (1.3)	1.0 (1.0)	1.0 (1.0)	13.0
$n=200$	1.2 (3.0)	1.0 (1.1)	1.0 (1.0)	1.0 (1.0)	9.0
$p=2$	$c=6$	$c=10$	$c=20$	$c=30$	$c(n,p)$
$n=25$	98.7 (411.8)	92.2 (239.5)	59.3 (145.2)	39.6 (55.1)	540.6
$n=50$	51.1 (238.5)	19.4 (53.0)	3.8 (7.3)	1.8 (3.2)	78.7
$n=100$	12.3 (66.4)	1.8 (4.5)	1.0 (1.1)	1.0 (1.0)	19.2
$n=200$	1.6 (8.6)	1.0 (1.3)	1.0 (1.0)	1.0 (1.0)	10.4
$p=3$	$c=6$	$c=10$	$c=20$	$c=30$	$c(n,p)$
$n=50$	195.1 (788.6)	179.7 (645.2)	103.0 (279.6)	18.5 (50.6)	199.9
$n=100$	79.3 (292.3)	9.1 (48.9)	1.2 (1.9)	1.0 (1.0)	35.2
$n=200$	4.6 (39.8)	1.0 (1.2)	1.0 (1.0)	1.0 (1.0)	13.0
$n=300$	1.3 (7.1)	1.0 (1.3)	1.0 (1.0)	1.0 (1.0)	10.1

4.2 Comparison of Methods

The performance of FPC analysis was compared to two other procedures from the literature by means of a Monte Carlo simulation, namely

Maximum Likelihood Clusterwise Linear Regression (MLCLR) as explained by DeSarbo and Cron (1988). They assume a fixed sequence of regressor values $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and model $y_1, \ldots, y_n$ as independently distributed according to

$$\mathcal{L}(y_i) = \sum_{j=1}^{k} \epsilon_j \mathcal{N}_{\mathbf{x}_i'\beta_j, \sigma_j^2},$$

where $\epsilon_j > 0$, $j = 1, \ldots, k$, and $\sum_{j=1}^{k} \epsilon_j = 1$. The ϵ_j denote the proportions of the k mixture components. They compute Maximum Likelihood estimators for the parameters $(\epsilon_j, \beta_j, \sigma_j^2)$, $j = 1, \ldots, k$, under fixed k by use of the EM-algorithm, which also provides estimators $\hat{\epsilon}_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, k$, for the probability that the point $(\mathbf{x}_i, y_i)$, conditional on its value, was generated by the mixture component j. The point $(\mathbf{x}_i, y_i)$ can then be classified as belonging to component

$j(i) := \arg\max_j\{\hat{\epsilon}_{ij}\}$. Wedel and DeSarbo (1995) propose the Consistent Akaike's Information Criterion (CAIC) of Bozdogan (1987) to estimate the number of mixture components k. I applied the algorithm with estimators from a random partition of the data points as starting values for the parameter estimators, an upper bound of 7 for k and a lower bound of 10^{-6} for the σ_j^2, $j = 1, \ldots, k$ (otherwise the likelihood function would be unbounded). The algorithm was terminated when the increase of the log-likelihood function fell below 10^{-7}. I implemented the method as described in the statistical programming language R.

Model Based Gaussian Clustering with Noise (MBGCN) as implemented in the software package MCLUST. A current version is treated in Fraley and Raftery (1998, 1999). They assume the points $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, as i.i.d. distributed according to

$$\mathcal{L}(\mathbf{z}_i) = \epsilon_0 \mathcal{U}_C + \sum_{j=1}^{k} \epsilon_j \mathcal{N}_{a_j, \mathbf{\Sigma}_j},$$

where $\mathcal{U}_C$ denotes the uniform distribution on a convex set C, $a_j \in I\!R^{p+1}$, $\mathbf{\Sigma}_j$ positive definite $(p+1) \times (p+1)$-covariance matrices for $j = 1, \ldots, k$, $\epsilon_j > 0$, $j = 0, \ldots, k$, and $\sum_{j=0}^{k} \epsilon_j = 1$. The $\mathbf{x}_i$-values from $I\!R^p$ do not include a component for the regression intercept in this setup. Such a Gaussian mixture model can also be applied to linear regression data, because a linear regression distribution is a $p+1$-variate Gaussian distribution if the distribution of the independent variables is assumed to be a p-variate Gaussian. DasGupta and Raftery (1998) propose the method for "highly linear" data. The mixture component $\mathcal{U}_C$ is designed to model noise or outliers not belonging to any of the Gaussian components. The covariance matrices $\mathbf{\Sigma}_j$ can be decomposed as $\mathbf{\Sigma}_j = \lambda_j \mathbf{D}_j \mathbf{A}_j \mathbf{D}_j$, where λ_j is the largest eigenvalue of $\mathbf{\Sigma}_j$, $\mathbf{D}_j$ is the matrix of eigenvectors, $\mathbf{A}_j = \mathrm{diag}(1, \alpha_{2j}, \ldots, \alpha_{(p+1)j})$. MCLUST is able to fit several models defined by various restrictions to λ, $\mathbf{A}$ and the $\mathbf{D}_j$-matrices. They are given in Table 1 of Fraley and Raftery (1999). The software computes Maximum Likelihood estimators using the EM-algorithm for the parameters $\epsilon_0, (\epsilon_1, a_1, \mathbf{\Sigma}_1), \ldots, (\epsilon_k, a_k, \mathbf{\Sigma}_k)$ from starting values computed by hierarchical clustering as explained in Fraley and Raftery (1999). The component memberships of the points can be estimated by analogy to the MLCLR procedure. The Bayesian Information Criterion BIC (Schwarz 1978) was used for the estimation of the number of components k as well as for the choice of an optimally restricted model.

An initial estimation of the noise component is needed. Fraley and Raftery (1998) propose the use of the software `NNclean` as explained in Byers and Raftery (1998). This software requires the choice of a constant K for the number of nearest neighbors of a point involved in the calculations. I used $K = 10$. $k \leq 7$ was again assumed. A lower bound for the covariance determinant (to bound the likelihood function) and a convergence criterion were used as implemented in MCLUST. The R-port version of MCLUST was used. The `Splus`-version of `NNclean`, which worked on R with only one small modification, was used.

Fixed Point Cluster analysis (FPCA) was carried out as described in Section 3.6.

Obviously the procedures differ with respect to their underlying models. MBGCN assumes Gaussian regressor distributions. The MLCLR model does not contain an outlier component, and it assumes the probability of $(\mathbf{x}_i, y_i)$ to be generated by mixture component j to be constant, regardless of $\mathbf{x}_i$. This assumption is called "assignment independence", see Hennig (2000b). It will be illustrated by the discussion of the simulated data constellations. The theory of FPCA (Hennig 2000c) is valid for any distribution on $I\!R^{p+1}$ but it is no exact estimation procedure for Gaussian mixture components. Instead, it estimates theoretical FPCs. Their existence is guaranteed corresponding to subpopulations of the type (1), where the rest of the data may be distributed arbitrarily, but well separated from the linear regression part. That is, I compare procedures that clearly do not estimate the same features of the data. However, all the methods may be applied to the same data with similar interpretations of the results. Therefore it is interesting to study the performance of their procedures in cases where the assumptions are not fulfilled, but where one may consider the methods as appropriate.

I have chosen four different constellations of n, p, the β, σ^2, G-parameters and noise for this paper. The simulations indicate that the results depend strongly upon all the parameter choices. An arbitrary "ranking" of the procedures could easily be illustrated by the choice of the appropriate constellation. I do not attempt to show that FPCA is generally better than the ML-methods, but at least there are some situations where it is superior.

The constellations are:

Square-p2: $n = 200$, $p = 2$, all x_1-values were generated by $\mathcal{U}_{[0,1]}$, $x_2 := x_1^2$. For the points 1-100: $y = x_1 + u$. For the points 101-200: $y = 0.5x_2 + u$, $\mathcal{L}(u) = \mathcal{N}_{0,0.0001}$ for all points. See Figure 2, left side. The assumptions of MLCLR are met because the data contain only linear regression clusters and the distribution of the independent variable does not vary be-

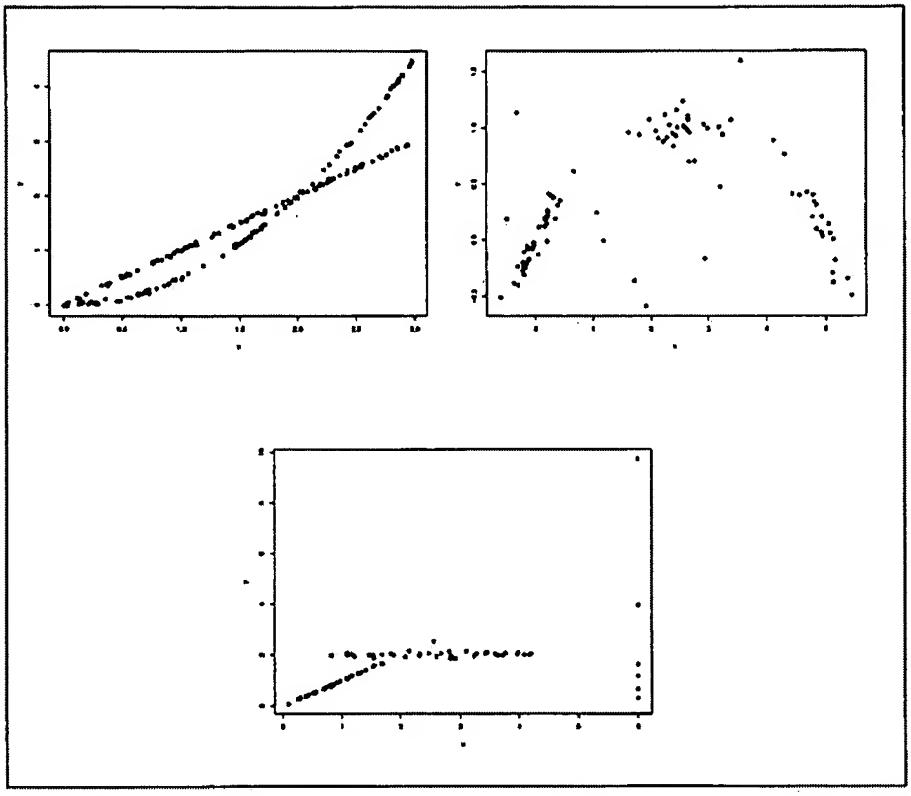

Figure 2. Typical data from the constellations "Square", "3+noise", and "2+noise".

tween clusters. The assumptions of MBGCN are not met because of the non-Gaussian regressor distribution.

Square-p1: The data sets of this constellation were generated as the data of the constellation `Square`, but the values of x_2 were not included, thus $n = 200$, $p = 1$. This is data with a linear and a nonlinear cluster. It meets only the assumptions of FPCA with respect to the remaining linear subpopulation (note that FPCA allows the rest of the data to be distributed arbitrarily).

3+Noise: $n = 100$, $p = 1$. Points 1-40 were generated by $\mathcal{L}(x_1) = \mathcal{N}_{0,0.09}$, $y = x_1 + u$, points 41-60 were generated by $\mathcal{L}(x_1) = \mathcal{N}_{5,0.09}$, $y = -x_1 + 5 + u$, points 61-90 were generated by $\mathcal{L}(x_1) = \mathcal{N}_{2.5,0.25}$, $y = 1 + u$, $\mathcal{L}(u) = \mathcal{N}_{0,0.01}$ for points 1-90. The points 91-100 were generated by $\mathcal{U}_{[-1,6]\times[-1,2]}$. See Figure 2, middle. The assumptions of MBGCN are met, approximately even that of an equal shape parameter A for all

Table 3: Average maximum similarity s_* of found cluster and average number of found clusters n_C for constellations "Square-p1" and "Square-p2"

Data	Square-p2			Square-p1		
Method	Pt. 1-100	101-200	n_C	1-100	101-200	n_C
MBGCN	0.461	0.457	5.67	0.856	0.616	4.10
MLCLR	0.983	0.979	2.10	0.973	0.652	3.93
FPCA	0.943	0.961	3.90	0.944	0.523	2.24

Table 4: Average maximum similarity s_* of nearest found cluster and average number of found clusters n_C for constellations "3+Noise" and "2+Noise"

Data	3+Noise				2+Noise		
Method	Pt. 1-40	41-60	61-90	n_C	1-40	41-75	n_C
MBGCN	0.990	0.980	0.973	3.05	0.942	0.969	2.55
MLCLR	0.694	0.714	0.529	3.04	0.934	0.914	3.68
FPCA	0.960	0.787	0.780	3.82	0.969	0.960	2.98

clusters. The assumptions of MLCLR are not met because of the noise and a strong violation of assignment independence: The domains of the regressors are nearly disjoint for the three clusters.

2+Noise: $n = 81$, $p = 1$. Points 1-40 were generated by $x_1 = |x_*|$, $\mathcal{L}(x_*) = \mathcal{N}_{0,1}$, $y = x_1 + u$, $\mathcal{L}(u) = \mathcal{N}_{0,0.0004}$. Points 41-75 were generated by $\mathcal{L}(x) = \mathcal{N}_{2.5,1}$, $y = 2 + u$, $\mathcal{L}(u) = \mathcal{N}_{0,0.0025}$. Points 71-73 were generated by $x = 6$, $y = 2 + u$, $\mathcal{L}(u) = \mathcal{N}_{0,4}$. See Figure 2, right side. This is a constellation of two regression clusters and six outliers. Neither the assumptions of MLCLR, nor those of MBGCN are met because the regressor distribution of the first cluster is not Gaussian, the noise is not uniform, and the regressor distributions of the two clusters differ.

There were 200 simulation runs for each method with each constellation. The results are given in the Tables 3 and 4. The estimated number of clusters n_C (meaning the number n_r of found representative FPCs in the case of FPC analysis) was recorded as well as the maximum similarity s_* between an estimated cluster and the given clusters of the constellation. n_C does not include the noise component in the case of MBGCN. The estimated number n_r of FPCs cannot be interpreted in the same manner as for the ML-methods, because

- FPCs may intersect or include each other (in particular there is frequently an FPC containing the whole dataset; in constellation "3+noise", sometimes the union of mixture components 2 and 3 forms an FPC, which makes some sense, see Figure 2, middle), and
- the number of found clusters of the other two methods was limited by seven.

Therefore, n_r can be expected to be slightly larger than the number of mixture components.

The average maximum similarity (over the simulation runs) was used as a measure of how good the methods discovered the given clusters. s_* was discussed in Section 3.4. I prefer this measure to mean (squared) errors of parameter estimators for two reasons:

- The methods aim to find the same clusters, but they do not estimate the same parameters.
- If a method fails essentially to find a cluster, it is not interesting if a parameter error is of size 1 or 1000, but such differences influence crucially the mean error of the simulation.

Where the model assumptions of one of the ML methods were fulfilled, the corresponding method led to the best results, as one should expect: MLCLR was best for "Square-p2", MBGCN was best for "3+Noise". In both cases, FPCA yielded better results than the misspecified ML-method. For the constellation "2+Noise", where the model assumptions of both ML methods were violated, FPCA led to the best results.

Both MBGCN and MLCLR did not mainly suffer from the outliers. MBGCN was more sensitive against strong non-normality of the independent variable in the constellations "Square", and MLCLR against the violation of the assignment independence in "3+Noise".

The performance of FPCA depends upon the size and the separateness of the cluster, as can be seen in "3+Noise". The second cluster, which is the smallest, and the third cluster, which has the largest error variance, are more difficult to find than a large, well separated cluster.

After subtracting one for the frequently occurring FPC of the whole dataset, the number n_r of FPCs can keep abreast of the ML methods as an estimator of the number of "cluster-shaped" mixture components. At "Square-p1", it is clearly the best.

5. Application to Tone Perception Data

The tone perception data stem from an experiment of Cohen (1984). A pure fundamental tone was played to a trained musician. Electronically generated overtones were added, determined by a stretching ratio of x. $x = 2.0$ corresponds to the harmonic pattern usually heard in traditional definite pitched instruments. The musician was asked to tune an adjustable tone to the octave above the fundamental tone. y gives the ratio of the adjusted tone to the fundamental, i.e. $y = 2.0$ would be the correct tuning for all x-values. The data analyzed here belong to 150 trials with the same musician. In the original study, there were four further musicians. The scatterplot suggests that there are

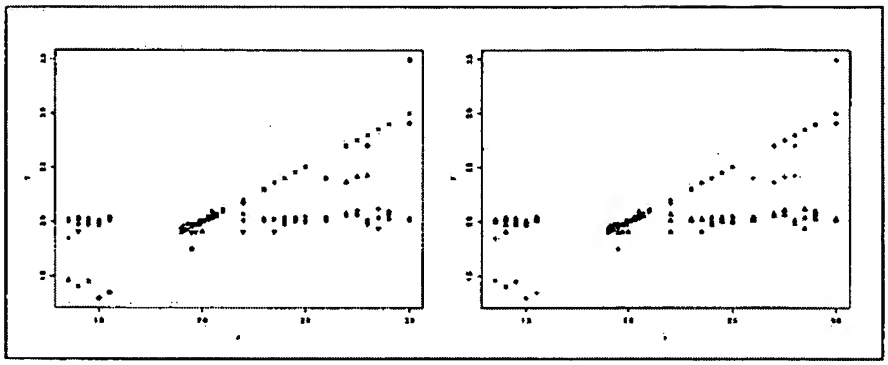

Figure 3. Partitions of the tone perception data from MLCLR and MBGCN.

two different linear regression populations, roughly corresponding to (somewhat biased) correct tuning and tuning to the first overtone, respectively, and some points that do not fit to any of them.

The application of MLCLR and MBGCN, as explained in Section 4, led to the partitions shown in Figure 3. MLCLR found six clusters. Among them were two clusters corresponding to correct tuning, and to overtone tuning, respectively. The occurrence of outliers not exactly consistent to the two most obvious lines caused the appearance of four further clusters. This is similar to the simulation results of "Square-p1" and "2+Noise", see Table 3 and 4, where MLCLR had been able to find the relevant clusters, but the number of clusters had frequently been overestimated in the presence of points not consistent with any linear regression line. A better fit might be possible with more (or better chosen) starting configurations for the EM-algorithm. De Veaux (1989) performed ML-estimation in the same model, assuming two mixture components instead of estimating their number. The result of MBGCN corresponds to the optical impression almost perfectly. There are the two main clusters and a noise component denoted by "+". This is a very useful result for a clusterwise linear regression, but it might be a little bit puzzling, however, from the viewpoint of the estimation of 2-dimensional Gaussian mixtures, because of the clear gap in the x-values between 1.6 and 1.9. The reason for points with large and small x-values to appear in the same clusters is that the distribution of the x-values for $x \geq 1.9$ is strongly right-skewed.

FPC analysis as described in Section 3.6 with $i_{n,p} = 853$, $c(n,p) = 10.07$ found twelve FPCs forming three Single Linkage groups. The three representative FPCs with 122, 74, and 66 points, are shown in Figure 5. Their parameters $(\beta_0, \beta_1, \sigma^2)$ are $(1.905, 0.048, 0.0028)$, $(0.023, 0.991, 0.0002)$, and $(0.794, 0.602, 0.0011)$. The numbers of times found (and expectation ratios

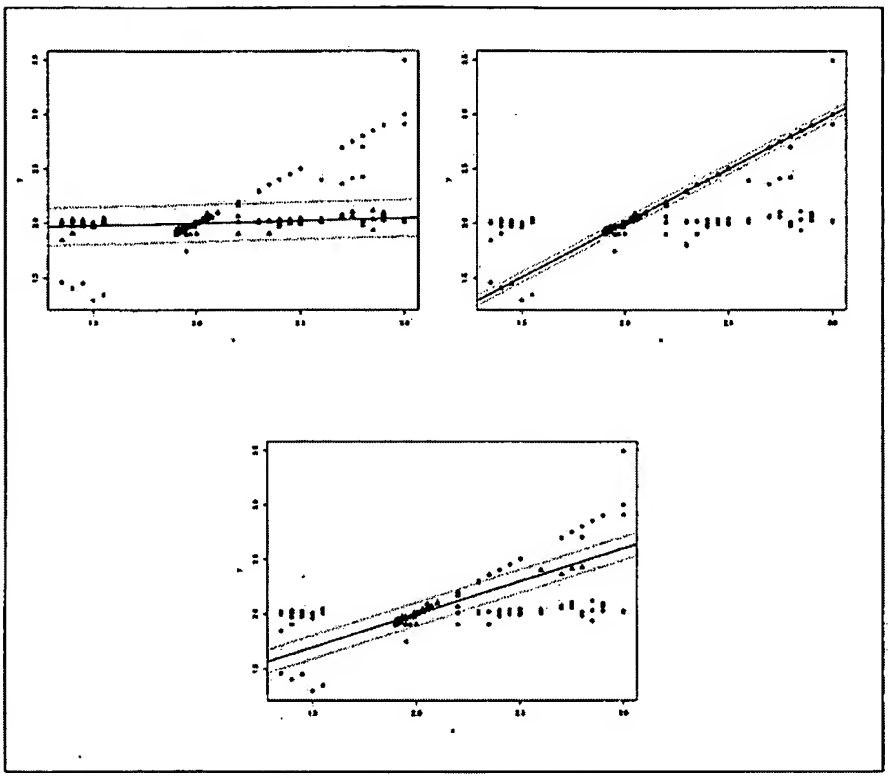

Figure 4. Representative FPCs of the tone perception data.

r_W) of the Single Linkage groups are 681 (1.49), 141 (1.41), and 10 (0.14). Seven points belong to none of the representative FPCs. The FPCs 1 and 2 correspond to the two clear lines (correct tuning and overtone tuning). Most of the points with $x \approx 2.0$ are consistent with both of the lines and are assigned to both clusters. The third FPC may appear somewhat surprising, but there is some justification for it, because the 59 points with $x \approx 2.0$ as a whole may be seen as a kind of compromise between correct and overtone tuning, and as such they are consistent with some points with larger x, which are not fitted well by the two main populations. However, the small value of $r_W = 0.14$ indicates a lower significance of this cluster compared to the others. A valid interpretation may be that the musician does not only tend always to hear the regular tone or the overtone as dominant, but sometimes, with $x > 1.9$ he is confused in such a way that he tunes his tone between the two possibilities. The points with approximately ordinary overtones ($x \approx 2.0$) cannot discriminate clearly between the three tendencies.

In conclusion, the FPCA solution gives a new insight to the data, which is useful if the aim of the analysis is exploratory, while it does not provide estimators for the parameters of a mixture model (especially not for the component proportions) because of the essential intersection of the clusters.

Note that the result of FPCA as well as that of MLCLR depends on chance because of the random selection of starting configurations. Applications of FPC analysis to other datasets (illustrating the same tendencies of the method) can be found in Hennig (1997, 1999, 2000c).

6. Conclusion

Fixed Point Cluster analysis was defined for clusterwise linear regression. An algorithm was developed, which aims to find the most significant clusters in the data with high probability and prevents meaningless FPCs. Some tuning constants are required. Their choice was discussed in detail. FPCA was compared with ML-estimators for a linear regression mixture model (MLCLR) and a model for $p + 1$-dimensional Gaussian mixtures (MBGCN/MCLUST) by means of a simulation study. Each ML-estimator performed best when its model assumptions were fulfilled. FPCA was usually better than an ML-estimator with strongly violated assumptions. Outliers cause MLCLR often to overestimate the number of clusters. Furthermore MLCLR suffers sometimes from violations of the assumption of independence of the assignment of points to mixture components of the x-values. MBGCN does not always lead to reasonable regression clusters if the independent variable deviates strongly from a Gaussian distribution. It does not appear to be very sensitive to non-uniform outliers.

The spirit of FPCA is more exploratory than that of ML-estimation of mixture model parameters. The ability of intersecting clusters allows for new insights in the data and enables FPCA not to be affected by identifiability problems of regression mixtures (Hennig, 2000b). A problem with the FPCA is the exponential increase of the computing time with the dimension p.

The principle of FPC analysis can be transferred to clustering problems apart from clusterwise linear regression, see Hennig (1998). The C-software `fixreg` can be obtained from:

http://www.math.uni-hamburg.de/home/hennig/

An `R/S-plus`-module is in preparation and will be available soon from the same web-site. The C-software needed about 8 seconds for a dataset from constellation "Square-p2" of Section 4.2 on a Sun UltraSPARC-IIi 333 MHz, while the R-function needed 22 minutes. The tone perception data were analyzed much faster (the R-function needed 3 minutes).

Appendix

Proof of Theorem 3.1. The proof is divided into three parts. Part 2 contains the main idea of the proof: Step 2 of the FPA should get the error variance $\hat{\sigma}^2(\mathbf{W}^k)$ small by adding points with squared residual $\leq c\hat{\sigma}^2(\mathbf{W}^k)$ and excluding those with larger residuals. Because $c > 1$, the inclusion of new points does not necessarily decrease the error variance. Part 2 of the proof will show that it can be multiplied by an appropriate factor $\pi_{n(\mathbf{W}^k)}$ such that the product $T = \pi_{n(\mathbf{W}^k)}\hat{\sigma}^2(\mathbf{W}^k)$ is always decreased. $\pi_{n(\mathbf{W}^k)}$ is a product of $n(\mathbf{W}^k)$ factors smaller than 1. That is, the FPA tries to unite many points ($\pi_{n(\mathbf{W}^k)}$ small) with small $\hat{\sigma}^2(\mathbf{W}^k)$. Part 1 shows $\hat{\beta}(\mathbf{W}^k)$ to be always uniquely defined.

Some notation:

$$\beta_k := \hat{\beta}(\mathbf{W}^k), \qquad \sigma_k^2 := \hat{\sigma}^2(\mathbf{W}^k),$$
$$M_k := (\mathbf{y} - \mathbf{X}\beta_k)'\mathbf{W}^k(\mathbf{y} - \mathbf{X}\beta_k) =$$
$$= (n(\mathbf{W}^k) - p - 1)\sigma_k^2, \qquad \pi_m := \prod_{i=p+1}^{m-1} p_i,$$
$$p_i := \left[1\left(\tfrac{c-1}{i-p} < 1\right)\left(1 - \tfrac{c-1}{i-p}\right) + 1\left(\tfrac{c-1}{i-p} \geq 1\right)\tfrac{1}{c}\right],$$
$$\mathbf{W}^+ := \max(\mathbf{W}^{k+1} - \mathbf{W}^k, 0), \qquad \mathbf{W}^- := \max(\mathbf{W}^k - \mathbf{W}^{k+1}, 0),$$

the maximum taken element-wise. $\mathbf{W}^+$ and $\mathbf{W}^-$ indicate the data points that are added, removed respectively, by step 2 of the FPA. Assume $\sigma_k^2 > 0$ for all k up to part 3.

Part 1: Show $\forall k: \; n(\mathbf{W}^k) > p + 1$ by complete induction, which means that β_k is uniquely defined $\forall k$. Recall $n(\mathbf{W}^0) > p + 1$ and show for $m \geq 0$: $n(\mathbf{W}^m) > p + 1 \Rightarrow n(\mathbf{W}^{m+1}) > p + 1$. By definition

$$|\{i : w_i^m = 1 \wedge (y_i - \mathbf{x}_i'\beta_m)^2 > c\sigma_m^2\}| = n(\mathbf{W}^-) \Rightarrow$$
$$\Rightarrow \sigma_m^2 \geq \tfrac{n(\mathbf{W}^-)c\sigma_m^2}{n(\mathbf{W}^m)-p-1} \Rightarrow n(\mathbf{W}^-) \leq \tfrac{n(\mathbf{W}^m)-p-1}{c}. \qquad (11)$$

Assuming $n(\mathbf{W}^m) \geq p + c + 1$:

$$n(\mathbf{W}^{m+1}) \geq n(\mathbf{W}^m) - n(\mathbf{W}^-) \geq \left(1 - \tfrac{1}{c}\right) n(\mathbf{W}^m) + \tfrac{p+1}{c} \geq$$
$$\geq \left(1 - \tfrac{1}{c}\right)(p + c + 1) + \tfrac{p+1}{c} = p + c > p + 1.$$

On the other hand $n(\mathbf{W}^-) \in I\!N$ and with (11):

$$n(\mathbf{W}^m) < p + c + 1 \Rightarrow 1 > n(\mathbf{W}^-) = 0 \Rightarrow n(\mathbf{W}^{m+1}) \geq n(\mathbf{W}^m).$$

Part 2: Show that

$$T : \mathrm{diag}(\{0,1\}^n) \mapsto [0, \infty) : \; \mathbf{W} \mapsto \pi_{n(\mathbf{W})}\sigma^2(\mathbf{W}) \qquad (12)$$

is strictly decreased by step 2 of the FPA unless $n(\mathbf{W}^+) = n(\mathbf{W}^-) = 0$, i.e., $\mathbf{W}^{k+1} = \mathbf{W}^k$. Therefore, no $\mathbf{W}^m$ with $\mathbf{W}^{m+1} \neq \mathbf{W}^m$ can be repeated during the FPA, and the FPA converges in a finite number of steps because there are only finitely many indicator vectors of length n.

Assume that $n(\mathbf{W}^+) > 0$, or $n(\mathbf{W}^-) > 0$ and show $T(\mathbf{W}^{k+1}) - T(\mathbf{W}^k) < 0$.

$$\begin{aligned} T(\mathbf{W}^{k+1}) - T(\mathbf{W}^k) &= \pi_{n(\mathbf{W}^{k+1})}\sigma_{k+1}^2 - \pi_{n(\mathbf{W}^k)}\sigma_k^2 = \\ &= \left(\frac{\pi_{n(\mathbf{W}^{k+1})}(n(\mathbf{W}^k)-p-1)}{n(\mathbf{W}^{k+1})-p-1} - \pi_{n(\mathbf{W}^k)}\right)\sigma_k^2 + \\ &\quad + \frac{\pi_{n(\mathbf{W}^{k+1})}}{n(\mathbf{W}^{k+1})-p-1}(M_{k+1} - M_k). \end{aligned} \tag{13}$$

Yield $M_{k+1} =$

$$\begin{aligned} &\min_{\beta}(\mathbf{y} - \mathbf{X}\beta_{k+1})'\mathbf{W}^{k+1}(\mathbf{y} - \mathbf{X}\beta_{k+1}) \leq \\ &\leq (\mathbf{y} - \mathbf{X}\beta_k)'\mathbf{W}^{k+1}(\mathbf{y} - \mathbf{X}\beta_k) \leq \\ &\leq M_k + n(\mathbf{W}^+)c\sigma_k^2 - n(\mathbf{W}^-)c\sigma_k^2 \end{aligned} \tag{14}$$

by definition of $\mathbf{W}^+$ and $\mathbf{W}^-$. If $n(\mathbf{W}^-) > 0$, there is strict "<". Hence with (13):

$$\begin{aligned} &T(\mathbf{W}^{k+1}) - T(\mathbf{W}^k) \leq \\ &\leq \left(\frac{\pi_{n(\mathbf{W}^{k+1})}(n(\mathbf{W}^k)-p-1+[n(\mathbf{W}^+)-n(\mathbf{W}^-)]c)}{n(\mathbf{W}^{k+1})-p-1} - \pi_{n(\mathbf{W}^k)}\right)\sigma_k^2 \end{aligned} \tag{15}$$

$=: q$. Let $d := n(\mathbf{W}^+) - n(\mathbf{W}^-) = n(\mathbf{W}^{k+1}) - n(\mathbf{W}^k)$. If $d = 0$ then

$$1 = \frac{\pi_{n(\mathbf{W}^k)}}{\pi_{n(\mathbf{W}^{k+1})}} = \frac{n(\mathbf{W}^k) - p - 1 + dc}{n(\mathbf{W}^{k+1}) - p - 1}$$

and therefore $q = 0$ and $T(\mathbf{W}^{k+1}) - T(\mathbf{W}^k) < 0$ (there is strict inequality in (15) because $d = 0$ and $\mathbf{W}^k \neq \mathbf{W}^{k+1}$ imply $n(\mathbf{W}^-) > 0$). Show further

$$d < 0 \Rightarrow \frac{\pi_{n(\mathbf{W}^k)}}{\pi_{n(\mathbf{W}^{k+1})}} \geq \frac{n(\mathbf{W}^k) - p - 1 + dc}{n(\mathbf{W}^{k+1}) - p - 1} \tag{16}$$

which implies $q \leq 0$ and again $T(\mathbf{W}^{k+1}) - T(\mathbf{W}^k) < 0$ by strict inequality in (15), and

$$d > 0 \Rightarrow \frac{\pi_{n(\mathbf{W}^k)}}{\pi_{n(\mathbf{W}^{k+1})}} > \frac{n(\mathbf{W}^k) - p - 1 + dc}{n(\mathbf{W}^{k+1}) - p - 1} \tag{17}$$

implying $q < 0$ and strict decrease of T.

Proof of (16): If $n(\mathbf{W}^{k+1}) \leq p + c$, then

$$\frac{n(\mathbf{W}^k) - p - 1 + dc}{n(\mathbf{W}^{k+1}) - p - 1} = 1 + \frac{(c-1)d}{n(\mathbf{W}^{k+1}) - p - 1} \leq 0 < \frac{\pi_{n(\mathbf{W}^k)}}{\pi_{n(\mathbf{W}^{k+1})}}.$$

On the other hand, with $n(\mathbf{W}^k) > n(\mathbf{W}^{k+1}) > p + c$, get

$$\frac{\pi_{n(\mathbf{W}^k)}}{\pi_{n(\mathbf{W}^{k+1})}} = \prod_{i=n(\mathbf{W}^{k+1})}^{n(\mathbf{W}^k)-1} \left(1 - \frac{c-1}{i-p}\right) \geq$$
$$\geq \left(1 - \frac{c-1}{n(\mathbf{W}^{k+1})-p-1}\right)^{-d}.$$

Use $(1-b)^m \geq 1 - bm$ for $0 < b < 1, m \in \mathbb{N}$. Let $b = \frac{c-1}{n(\mathbf{W}^{k+1})-p-1}$, $m = -d$. Then

$$\frac{\pi_{n(\mathbf{W}^k)}}{\pi_{n(\mathbf{W}^{k+1})}} \geq 1 + d\frac{c-1}{n(\mathbf{W}^{k+1}) - p - 1} = \frac{n(\mathbf{W}^k) - p - 1 + dc}{n(\mathbf{W}^{k+1}) - p - 1}.$$

Proof of (17): By complete induction over $m > 0$ get for $b > 0, m \in \mathbb{N}$:

$$1 - \frac{(c-1)m}{b+mc} > \left(1 - \frac{c-1}{b+m}\right)^{m_1} \left(\frac{1}{c}\right)^{m_2}$$
$$\forall m_1, m_2 \in \mathbb{N}_0 \text{ where } m_1 + m_2 = m \tag{18}$$
$$\text{assuming } c > 1 \text{ and } \frac{c-1}{b+m} < 1. \tag{19}$$
$$\text{In particular } 1 - \frac{(c-1)m}{b+mc} = \frac{b+m}{b+mc} > \frac{1}{c} \geq \left(\frac{1}{c}\right)^m. \tag{20}$$

Start with $n(\mathbf{W}^{k+1}) > p + c - 1$. Because $n(\mathbf{W}^{k+1}) > n(\mathbf{W}^k) > p + 1$, apply (18) with $m = d$, $m_1 = n(\mathbf{W}^{k+1}) - \max(n(\mathbf{W}^k), \lfloor p + c - 1 \rfloor + 1)$, $b = n(\mathbf{W}^k) - p - 1$. (19) is fulfilled because $\frac{c-1}{b+m} = \frac{c-1}{n(\mathbf{W}^{k+1})-p-1} < 1$.

$$\frac{\pi_{n(\mathbf{W}^{k+1})}}{\pi_{n(\mathbf{W}^k)}} = \prod_{i=n(\mathbf{W}^k)}^{n(\mathbf{W}^{k+1})-1} p_i =$$

$$= \prod_{\max(n(\mathbf{W}^k), \lfloor p+c-1 \rfloor + 1) < i \leq n(\mathbf{W}^{k+1})-1} \left(1 - \frac{c-1}{i-p}\right) \prod_{i=n(\mathbf{W}^k)}^{\lfloor p+c-1 \rfloor} \frac{1}{c} \leq$$
$$\leq \left(1 - \frac{c-1}{n(\mathbf{W}^{k+1})-p-1}\right)^{m_1} \left(\frac{1}{c}\right)^{d-m_1} <$$
$$< 1 - \frac{(c-1)d}{n(\mathbf{W}^k)-p-1+dc} = \frac{n(\mathbf{W}^{k+1})-p-1}{n(\mathbf{W}^k)-p-1+dc},$$

i.e., (17). With $n(\mathbf{W}^{k+1}) \leq p + c - 1$ get

$$\frac{\pi_{n(\mathbf{W}^{k+1})}}{\pi_{n(\mathbf{W}^k)} = \left(\frac{1}{c}\right)^d < \frac{n(\mathbf{W}^{k+1})-p-1}{n(\mathbf{W}^k)-p-1+dc}}$$

because of (20).

Part 3: Assume $\sigma_k^2 = 0$ for some k. $\sigma_{k-1}^2 > 0 \Rightarrow n(\mathbf{W}^k) > p+1$ because of part 1. Further, $\mathbf{W}_i^k = 1 \Rightarrow (y_i - \mathbf{x}_i'\beta_k)^2 = 0, w_i^{k+1} = 1 \Leftrightarrow (y_i - \mathbf{x}_i'\beta_k)^2 = 0$ and $n(\mathbf{W}^{k+1}) \geq n(\mathbf{W}^k) > p+1$. Hence $\sigma_{k+1}^2 = \sigma_k^2 = 0$, $\beta_{k+1} = \beta_k$, $\mathbf{W}^{k+2} = \mathbf{W}^{k+1}$.

References

BOZDOGAN, H. (1987), "Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions", *Psychometrika, 52*, 345-370.

BYERS, S., and RAFTERY, A. E. (1998), "Nearest Neighbor Clutter Removal for Estimating Features in Spatial Point Processes," *Journal of the American Statistical Association, 93*, 577-584.

COHEN, E. (1984), "Some effects of inharmonic partials on interval perception", *Music Perception, 1*, 323-349.

COLEMAN, D. A. and WOODRUFF, D. L. (2000), "Cluster Analysis for Large Datasets: An Effective Algorithm for Maximizing the Mixture Likelihood", *Journal of Computational and Graphical Statistics, 9*, 672-688.

COOK, R. D. and CRITCHLEY, F. (2000), "Identifying Regression Outliers and Mixtures Graphically," *Journal of the American Statistical Association, 95*, 781-794.

DASGUPTA, A. and RAFTERY, A. E. (1998), "Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering", *Journal of the American Statistical Association, 93*, 294-302.

DAVIES, P. L. and GATHER, U. (1993), "The Identification of Multiple Outliers," with discussion, *Journal of the American Statistical Association, 88*, 782-801.

DE VEAUX, R. D. (1989), "Mixtures of Linear Regressions", *Computational Statistics and Data Analysis, 8*, 227-245.

DESARBO, W. S., and CRON, W. L. (1988), "A maximum likelihood methodology for clusterwise linear regression", *Journal of Classification, 5*, 249-282.

FRALEY, C., and RAFTERY, A. E. (1998), "How Many Clusters? Which Clustering Method? Answers Via Model Based Cluster Analysis," *Computer Journal, 41*, 578-588.

FRALEY, C., and RAFTERY, A. E. (1999), "MCLUST: Software for Model-Based Cluster Analysis," *Journal of Classification, 16*, 297-306.

GARCIA-ESCUDERO, L. A., and GORDALIZA, A. (1999), "Robustness Properties of k Means and Trimmed k Means", *Journal of the American Statistical Association, 94*, 956-969.

HAMPEL, F. R. (1975), "Beyond location parameters: Robust concepts and methods", *Bulletin of the International Statistical Institute 46 - Proceedings of the 40th Session*, 375-382.

HAWKINS, D. M. (1999), "Improved feasible solution algorithms for high breakdown estimation", *Computational Statistics and Data Analysis, 29*, 145-161.

HENNIG, C. (1997), "Fixed Point Clusters and their Relation to Stochastic Models" in *Classification and Knowledge Organization*, eds. Klar, R. and Opitz, O., Berlin: Springer-Verlag, 20-28.

HENNIG, C. (1998), "Clustering and Outlier Identification: Fixed Point Cluster Analysis" in *Advances in Data Science and Classification*, eds. A. Rizzi, M. Vichi, and H.-H. Bock, Berlin: Springer-Verlag, 37-42.

HENNIG, C. (1999), "Models and Methods for Clusterwise Linear Regression", in *Classification in the Information Age*, ed. W. Gaul, and H. Locarek-Junge, Heidelberg: Springer-Verlag, 179-187.

HENNIG, C. (2000a), "What Clusters are Generated by Normal Mixtures?" in *Data Analysis, Classification and Related Methods*, eds. H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen, M. Schader, Berlin: Springer-Verlag, 53-58.

HENNIG, C. (2000b), "Identifiability of Models for Clusterwise Linear Regression", *Journal of Classification, 17*, 273-296.

HENNIG, C. (2000c), "Clusters, Outliers and Regression: Fixed Point Clusters", to appear in *Journal of Multivariate Analysis*.

HUBER, P. J. (1981), *Robust Statistics*, New York: Wiley.

KHARIN, Y. (1996), *Robustness in Statistical Pattern Recognition*, Dordrecht: Kluwer Academic Publishers

KOSINSKI, A. S. (1999), "A procedure for the detection of multivariate outliers", *Computational Statistics and Data Analysis, 29*, 145-161.

MCLACHLAN, G. J., PEEL, D., BASFORD, K. E., and ADAMS, P. (1999), "The EMMIX software for the fitting of mixtures of normal and t-components", *Journal of Statistical Software, 4*, No. 2.

MORGENTHALER, S. (1990), "Fitting Redescending M-Estimators in Regression," in *Robust Regression*, eds. H. D. Lawrence, and S. Arthur, New York: Dekker, 105-128.

PEEL, D. and MCLACHLAN, G. J. (2000), "Robust mixture modelling using the t-distribution", *Statistics and Computing 10*, 335-344.

ROCKE, D. M. and WOODRUFF, D. L. (1996), "Identification of Outliers in Multivariate Data", *Journal of the American Statistical Association, 91*, 1047-1061.

ROCKE, D. M. and WOODRUFF, D. L. (2001), Discussion on Pena, D. and Prieto, F. J. "Multivariate Outlier Detection and Robust Covariance Matrix Estimation", *Technometrics, 43*, 300-303.

ROUSSEEUW, P. J. (1994), "Unconventional features of positive-breakdown estimators", *Statistics and Probability Letters, 19*, 417-431.

ROUSSEEUW, P. J. and LEROY, A. M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

SCHWARZ, G. (1978), "Estimating the dimension of a model", *Annals of Statistics, 6*, 461-464.

SHANNON, W. D., FAIFER, M., PROVINCE, M. A. and RAO, D. C. (2002), "Tree-Based Models for Fitting Stratified Linear Regression Models", *Journal of Classification, 19*, 113-130.

VIELE, K. and TONG, B. (2000), *Modeling with Mixtures of Linear Regression*, Technical Report, University of Kentucky.

WEDEL, M. and DESARBO, W. S. (1995), "A mixture likelihood approach for generalized Linear models", *Journal of Classification, 12*, 21-56.

WEDEL, M. and KAMAKURA, W. (2000), *Market Segmentation* (2nd ed.), Boston: Kluwer Academic Publishers.